bio-protocol Preprint

Not Peer Reviewed
Ask a question online

Cell state transition inference

Liangtao Zheng Shishang Qin Xueda Hu Zemin Zhang

Updated date: Jul 23, 2022

An abbreviated version of this protocol was published in Science in Dec 2021

Pan-cancer single-cell landscape of tumor-infiltrating T cells

DOI: 10.1126/science.abe6474

Detailed protocol

Cell state transition inference

1) T cell differentiation trajectory inference

To model the T cell state transition, the diffusion map algorithm *(25)*, which preserves the global relations and pseudotemporal ordering of cells, was applied to infer the differentiation trajectory. We fed the mini-cluster expression matrix and the previously calculated principal components matrix into the scanpy pipeline (version 1.6.1). A neighborhood graph based on principal components was constructed using the *scanpy.pp.neighbors* function. Given the neighborhood graph, the diffusion map was built using *scanpy.tl.diffmap* function. The first two diffusion components (DCs) were used for visualization. Depending on the positions of less differentiated cells (e.g. naïve $CD8^+$ T cells) on the diffusion map, the cell with a maximum (when the less differentiated cells were on the right of the map)/minimum (when the less differentiated cells were on the left of the map) value in the DC1 was specified as the root cell, then the diffusion pseudotime *(58)* was calculated using *scanpy.tl.dpt* function. The original diffusion pseudotime values were further converted to percentile rank values as described in *(27)*. To find the potential genes driving the differentiation process, we fit a generalized additive model (*gam* function in the gam package of R) for the pseudotime and the expression profile of each gene. Genes with absolute coefficient > 0.5 and FDR < 0.01 were considered as the dynamic genes. The dynamic genes were then clustered based on the cosine distance. To find the pseudotime zone where a gene had a high acceleration in expression, indicating that a switch-like change may occur *(58)*, we calculated the first derivative of the fitted line, then the zone was defined as the range from the earliest pseudotime, where its first derivative passed 1, to the pseudotime, where its first derivative decreased below 1 for the first time.

To confirm the trajectory result from the diffusion map, Monocle3 *(27)* was applied to the mini-cluster expression matrix and the UMAP embedding, which can better preserve the local relations of cells. Specifically, the principal components were re-computed from the mini-cluster expression matrix. Based on the function *cluster_cells*, the mini-clusters were divided into large and well-separated groups called partitions, within each of which a principal graph was fitted using the function *learn_graph*. The principal graph was shown on the UMAP as "skeleton lines", indicating the differentiation trajectories. Assigning mini-clusters to the nearest principal graph nodes, the principal graph node containing the highest fraction of naive T cells was specified as the root, then the pseudotime was calculated using function *order_cells*.

2) RNA velocity analysis

The RNA velocity analyses were performed on the newly generated 10x scRNA-Seq data. The spliced and unspliced UMIs for each gene in each cell were counted using the python package velocyto (version 0.17.15). The subsequent analyses were performed by scanpy (version 1.6.1) and scvelo *(26)*. Generally, the count matrices were first normalized by library size and then filtered to keep only the genes that could be detected in > 30 cells commonly for both spliced and unspliced matrices, and the genes exhibiting high variability. Then a k nearest-neighbor graph (k=30) was built using the top 30 principal components, which were from a PCA performed upon the logarithmic spliced matrix with cell cycle score regressed out. For each cell, the moments (means and uncentered variances) of normalized spliced/unspliced counts were computed using the 30 nearest neighbors by function *scv.pp.moments*. The moments facilitated the RNA velocity estimation implemented in function *scv.tl.velocity* with mode set to "dynamical". The estimated velocities were used to construct a velocity graph representing the transition probabilities among cells by function *scvelo.tl.velocity_graph*. Finally, the velocity graph was used to embed the RNA velocities into the UMAP or diffusion map on a grid by the function *scv.pl.velocity_embedding_grid*.

In the global analyses of $CD8^+$ T cells and $CD4^+$ T cells, and the state transition analyses of $CD4^+$Tfh related populations and $CD4^+$ Treg populations, to reduce computational consumption, we applied mini-cluster level analyses, implemented by averaging the normalized counts per mini-cluster. It should be noted that gene selection impacted the result of the RNA velocity and the explanation. In the global analyses of $CD8^+$ T cells and $CD4^+$ T cells, we set the highly variable genes to the informative genes which were involved in the meta-cluster analysis (see section "Data integration and meta-cluster identification"). Those genes should be more powerful in distinguishing different cell states than the genes identified in an unsupervised manner such as the default implementation of scvelo. In the analysis of exhaustion, we set the highly variable genes to the informative genes among the meta-clusters of the two major exhaustion paths. We defined the “exhaustion program” as the top 50 signature genes of terminal Tex. To visualize the transition potential of the “exhaustion program”, those genes were embedded into the UMAP. In other cases, the RNA velocities of the highly variable genes were embedded into the UMAP or diffusion map, and should be explained as the transition potential of “the overall transcriptomic state”.

3) Inference of the two major exhaustion paths of $CD8^+$ T cells

To investigate the exhaustion process, we first performed diffusion map analysis using the informative genes. The result clearly showed that naive T cells were in one branch, and Temra and Tex were in two different branches (fig. S13, A and B), indicating that T cells could differentiate into two major different fates. More importantly, the pseudotime ordering shows that *GZMK*$^+$ meta-clusters and *ZNF683*$^+$*CXCR6*$^+$Trm were located between naive T cells and terminal Tex, especially most of *GZMK*$^+$Tem, *GZMK*$^+$Tex, and *ZNF683*$^+$*CXCR6*$^+$Trm cells were located in the late stage of the pseudotime (>50%) (fig. S13C). These observations suggested that the *GZMK*$^+$ T cells and *ZNF683*$^+$Trm cells were pre-exhausted T cells, and globally they were not in the differentiation terminals. The Monocle 3 and UMAP analysis using the same informative genes showed similar results (fig. S13, D and E). Since UMAP

could better preserve the local relations of cells *(28)*, two pre-exhaustion trajectories could be distinguished in the UMAP and Monocle 3 result. To better present the data, we re-generated a UMAP using only the meta-clusters in the two major exhaustion paths and re-computed informative genes. Visually, it is apparent that there were two distinctive paths in the UMAP. The two paths differed in the expression of genes, including *GZMK* and *ZNF683*, which exhibited a mutually exclusive expression pattern (fig. S14A). Accordingly, the expression of the exhaustion program increased significantly along the trajectories (fig. S14B). Similarly, when mapping the RNA velocities of genes from the exhaustion program into the UMAP, it was clear that the transition direction of the exhaustion program was from the pre-exhaustion states to exhaustion (fig. S14C). Thus, the gene expression-based analyses suggested two exhaustion paths of $CD8^+$ T cells.

We not only used gene expression data but also used TCRs as markers to trace the state transitions. We found many shared TCRs between terminal Tex and *GZMK*$^+$Tex or *ZNF683*$^+$*CXCR6*$^+$Trm cells. Those TCR sharings were summarized as pTrans (Fig. 2C). The TCR sharings between different states should be solid evidence supporting the differentiation relationships. Based on this, we inferred the T cell state transition paths. Specifically, the pTrans suggested developmental connections among meta-clusters and formed a graph. As the naive T cells (CD8.c01) had few TCR sharing with other T cells, and the nearest meta-cluster neighbor of CD8.c01 was CD8.c02 based on UMAP, therefore it was reasonable to add the developmental connection between them to the graph. The edges of the graph could be assigned a "cost", i.e. the Euclidean distance between the median centers of the meta-clusters in the UMAP. This graph could be reduced to infer the T cell state transition path under certain criteria. We considered the following criteria: 1) naive T cells (CD8.c01) was the start point of the reduced paths, and the terminal exhausted T cells (CD8.c12) was the endpoint of the reduced paths; 2) every meta-cluster must be in at least one path; 3) the sum of the cost of all possible paths beginning from the start point to the endpoint in the reduced graph was minimum. Under the constraints of the criteria, the graph was simplified to a two-path graph, which contained the two exhaustion paths (fig. S14D).

For comparison, the algorithm of PAGA *(59)* was applied, and it identified a graph that also contained two paths (fig. S14D). The graph identified by PAGA shares the "*GZMK*$^+$ path" with the STARTRAC-based path inference described above, but the "Trm path" was more "expensive" (higher total length of the path). Slingshot *(60)* was also applied, which was designed to identify a minimum spanning tree (MST), although the MST structure might not apply to all data *(61)*. In this particular setting, in the inferred paths by Slingshot (fig. S14D), CD8.c05(*GZMK*$^+$ early Tem) was classified as a terminal cell state, which conflicts with the diffusion map results. To quantitatively compare the goodness of fitting using different graphs, we calculate a ratio of the sum of squares between two graphs. Specifically, in the UMAP, the Euclidean distances of one data point to the segments of graph G were calculated. Then the minimum distance of the data point to the graph was defined as the minimum of those Euclidean distances. The sum of squares of the minimum distances of all data points to G (SS_G) measured the deviation of the data points from the fitted graph. Given two graphs A and B, the ratio SS_A/SS_B was used to evaluate which graph fitted the data points better. The value of ratio SS_A/SS_B less than 1 indicated the deviation of the fitting by B was higher than that by A, and hence graph A fitted the data points better. The ratio SS_A/SS_B was far less than 1, the fitting by A much better than the fitting by B. In this study, we set graph A to the graph we inferred, and graph B to one graph from other methods, including PAGA and Slingshot. Based on the SS_A/SS_B measurement, our inferred two-path model had higher goodness of fitting than the PAGA, Slingshot, and one-path model (fig. S14E).

Taken together, with the evidence from gene expression-based trajectory analysis (i.e. globally *GZMK*$^+$ T cells and *ZNF683*$^+$Trm cells were in earlier pseudotime, they had lower expression of exhaustion programs, and occupied the upstream positions in the RNA velocity field), and with the evidence from TCR sharings, we made the suggestion that T cells followed the two major paths to exhaustion, through Tem and Trm respectively.

4) Assessment of the enrichment of the two major exhaustion paths in individual clonotypes

To assess the enrichment of the two major exhaustion paths in individual clonotypes, we performed permutation tests using the expanded clonotypes containing both terminal Tex and cells of P1-specific populations (*GZMK*$^+$ early Tem, *GZMK*$^+$ Tem, and *GZMK*$^+$ Tex) or P2-specific population (*ZNF683*$^+$*CXCR6*$^+$ Trm). Specifically, for each clonotype, a ratio of the cell number in P1 populations to that in P1 and P2 populations was calculated. The null distribution of such a ratio could be obtained by random permutation of the cluster labels 1,000 times. The parameters of the normal-like null distribution were estimated by the *fitdistr* function in the R package MASS, then the two-side *p*-value for the actual ratio could be obtained. Clonotypes with adjusted *p*-values by the FDR method < 0.05 were considered as showing significant enrichment of P1 or P2 differentiation.

5) Inference of state transition between ISG$^+$ CD8$^+$ T cells and exhaustion

First, ISG$^+$CD8$^+$ T cells were a mixture of Tem, Trm, or even terminal Tex, and not a differentiation terminal but an intermediate state. To demonstrate this, we mapped the ISG$^+$ CD8$^+$ T cells to the two major exhaustion trajectories using a weighted-nearest neighbor (WNN)-based reference mapping algorithm implemented in Seurat (version 4). Specifically, the CD8$^+$ ISG$^+$ T (CD8.c15) meta-cluster was used as the query, and meta-clusters in the two major exhaustion paths were used as the reference. Genes used for this analysis were the same as those used for generating the UMAP of the reference. We identified anchors between the reference and the query in PCA space using the function *FindTransferAnchors*. Then the reference labels were transferred to the query using the function *TransferData*. We kept only mini-clusters with prediction scores > 0.5. Then those mini-clusters were mapped to the reference UMAP using the functions *IntegrateEmbeddings* and *ProjectUMAP* sequentially. We found that most of the ISG$^+$CD8$^+$ T cells could be mapped to meta-clusters in the two major CD8$^+$ exhaustion paths. Thus, the ISG$^+$ state was not an independent and terminal state, but a mixture of Tem, Trm and other cells (fig. S17).

Since ISG$^+$ CD8$^+$ T cells also had many TCR sharings with terminal Tex, we inferred that terminal Tex could differentiate from ISG$^+$ T cells. This claim is consistent with the previous studies where the ISG$^+$ state was an intermediate state in human T cell activation *(29)*, and the ISG$^+$ state was prone to a terminal state *(30,31)*. We also identified interferons as potential ligands inducing the transcription programs of Tex (fig. S35D), and identified high expression of interferon-related transcription factors in Tex such as *IFI16*, *IRF5*, *ETV1,* and *ETV7* (Fig. 2, H and I, table S3). Taken together, we suggested that interferon impacts the exhaustion process of tumor-infiltrating T cells.

6) Identification of transcription factors with differential usage between the two major exhaustion paths

The existence of transcription factors (TFs) differentially expressed along the two major exhaustion paths could be another supporting evidence for the existence of two different major exhaustion paths. To identify such TFs, we first identified four cell populations at transitional stages with the following criteria. Cell population in early stage of P1: (1) meta-cluster was one of CD8.c01(Tn), CD8.c02(*IL7R*$^+$ Tm), CD8.c05(*GZMK*$^+$ early Tem), and CD8.c06(*GZMK*$^+$ Tem); (2) diffusion pseudotime percentile <0.4; (3) median point of CD8.c01(Tn) < UMAP1 < median point of CD8.c06(*GZMK*$^+$ Tem); (4) UMAP2 < median point of CD8.c01(Tn). Cell population in late stage of P1: (1) meta-cluster was one of CD8.c06(*GZMK*$^+$ Tem), CD8.c11(*GZMK*$^+$ Tex), and CD8.c12(terminal Tex); (2) diffusion pseudotime percentile >0.6; (3) median point of CD8.c06(*GZMK*$^+$ Tem) < UMAP1 < median point of CD8.c12(terminal Tex); (4) UMAP2 < median point of CD8.c12(terminal Tex). Cell population in early stage of P2: (1) meta-cluster was one of CD8.c01(Tn), CD8.c02(*IL7R*$^+$ Tm), and CD8.c10(*ZNF683*$^+$*CXCR6*$^+$ Trm); (2) diffusion pseudotime percentile <0.4; (3) median point of CD8.c01(Tn) < UMAP1 < lower quartile of CD8.c10(*ZNF683*$^+$*CXCR6*$^+$ Trm); (4) UMAP2 > median point of CD8.c01(Tn). Cell population in late stage of P2: (1) meta-cluster was one of CD8.c10(*ZNF683*$^+$*CXCR6*$^+$ Trm), and CD8.c12(terminal Tex); (2) diffusion pseudotime percentile >0.6; (3) upper quartile of CD8.c06(*GZMK*$^+$ Tem) < UMAP1 < median point of CD8.c12(terminal Tex); (4) UMAP2 > median point of CD8.c12(terminal Tex). Then, in each population, we fit a generalized additive model (*gam* function in the gam package of R) for the pseudotime and the expression profile of each TF. TFs with absolute coefficients of the

1. Zheng, L. , Qin, S. , Hu, X. and Zhang, Z. (2022). Cell state transition inference. Bio-protocol Preprint. bio-protocol.org/prep1814.
2. Zheng, L., Qin, S., Si, W., Wang, A., Xing, B., Gao, R., Ren, X., Wang, L., Wu, X., Zhang, J., Wu, N., Zhang, N., Zheng, H., Ouyang, H., Chen, K., Bu, Z., Hu, X., Ji, J. and Zhang, Z.(2021). Pan-cancer single-cell landscape of tumor-infiltrating T cells. Science 374(6574). DOI: 10.1126/science.abe6474